EVALUATION OF A SPOKEN DIALOGUE SYSTEM FOR VIRTUAL REALITY CALL FOR FIRE TRAINING

Susan M. Robinson, Antonio Roque, Ashish Vaswani and David Traum*
Institute for Creative Technologies, University of Souhern California
13274 Fiji Way, Marina del Rey, CA, 90292

Charles Hernandez
Army Research Labs, HRED Field Element
Fort Sill, Lawton, OK

Bill Millspaugh
Tec-Masters, Inc., Lawton, OK

ABSTRACT

We present an evaluation of a spoken dialogue system that engages in dialogues with soldiers training in an immersive Call for Fire (CFF) simulation. We briefly describe aspects of the Joint Fires and Effects Trainer System, and the Radiobot-CFF dialogue system, which can engage in voice communications with a trainee in call for fire dialogues. An experiment is described to judge performance of the Radiobot CFF system compared with human radio operators. Results show that while the current version of the system is not quite at human-performance levels, it is already viable for training interaction and as an operator-controller aid.

1. INTRODUCTION

Radiobots are spoken dialogue systems that communicate over the radio in support of military training simulations. In this paper we describe the design and results of the evaluation of the first version of our Radiobot-CFF system (Roque et al, 2006b). Radiobot-CFF receives spoken radio calls for artillery fire from a forward observer team in a simulation-based training environment, and is able to carry on the Fire Direction Center (FDC) side of a conversation with the observer, while sending appropriate messages to a simulator to engage in the requested missions. Radiobot-CFF has been integrated with Firesim XX1[1] and the Urban Terrain Module (UTM) of the Joint Fires and Effects Trainer System (JFETS) of Fort Sill, Oklahoma.

Current training in the UTM often involves multiple simulation operators to engage with a single observer team: one operator to act as fire support officer (FSO) and talk with the observer team on the radio, and one to deal with technical aspects of the FDC, filling in information and monitoring a simulation GUI of students.[2] One of the goals of the **Radiobots-CFF** project was to provide spoken language technology to increase both the efficiency and effectiveness of the training process by automating the bulk of the FDC tasks, allowing a single operator to monitor and instruct students. Radiobot-CFF can be run in 3 different modes, depending on the level of support and direct engagement an operator would like to take. In automatic mode, the Radiobot can handle all communications with the simulator and trainees, without any operator intervention. In semi-autonomous mode, the observer must verify the suggested moves of the radiobot, and has an opportunity to change the understanding or course of actions. Finally, in manual mode, the radiobot simply observes the interaction, providing a transcript of its understanding for later review. An operator is also free to change modes during the course of the dialogue. While we have not yet had a chance to test it, use of Radiobot-CFF would also make it possible to conduct multiple missions with multiple FO teams per instructor, thus increasing the cost-effectiveness and rate of training of operator involvement for a large group of trainees.

The evaluation of the Radiobot-CFF system was conducted over several sessions on site with a total of 63 soldiers from the Field Artillery School at Fort Sill.

The rest of this paper is organized as follows: In section 2 we describe the Radiobots-CFF domain and JFETS UTM traininer in more detail. In section 3, we

[1] http://sill-www.army.mil/blab/sims/FireSimXXI.htm

[2] It is possible for both roles to be played by a single operator/controller, though this requires greater attention to simulator mechanics and leaves even less ability for focusing on learning objectives of trainees.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluation of a Spoken Dialogue System for Virtual Reality Call for Fire Training	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Southern California,Institute for Creative Technologies,13274 Fiji Way,Marina del Ray,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

describe the Radiobot-CFF system. In section 4, we describe the evaluation methodology and metrics used. Section 5 includes description of the evaluation experiments at Ft Sill, and results are given in Section 6. We conclude in section 7 with some analysis and future directions.

2. CALL FOR FIRE TRAINING

The JFETS UTM is a training environment with the objective of training U.S. army soldiers in the procedures of calls for artillery fire by practicing in a realistic urban environment. The UTM is fully immersive: in the course of a session, Fire Support (FS) Officers and Soldiers enter a room built to resemble an apartment in the Middle East, with a window view of a city below, as shown in figure 1.

Figure 1 UTM training environment

The city view is a rear-projected computer display. FS students view close-ups of the city and acquire targets through binoculars that have been modified to synchronize with the graphics display. Calls for fire are made via radio to one or more instructors or operators, who play the role of a fire direction center (FDC) in a room below. The operator enters mission information into a control panel, which results in the generation of a fire mission and the simulated effects (both graphic and audio) of the fires. Ambient sounds of the city are also audible throughout the session, and climate controls in the room approximate that of the Middle East.

Calls for fire follow a procedure outlined in an army tactics, techniques, and procedures manual (Department of the Army, 1991). When the forward observer has located a target, he conveys the location and target details to his team member, the RTO, who then initiates a call for fire. A fire mission follows a fairly strict procedure; a typical example is shown in figure 2.

A CFF can be roughly divided into three phases. In the first phase (utterances 1-6 of figure 2), the RTO identifies himself and the type of fire he is requesting (line 1), the target coordinates (line 3), and the target description and type of rounds requested (line 5). In this phase, the FSO simply repeats and confirms each bit of information.

1	RTO	steel one niner this is gator niner one adjust fire polar over
2	FSO	gator nine one this is steel one nine adjust fire polar out
3	RTO	direction five niner four zero distance four eight zero over
4	FSO	direction five nine four zero distance four eight zero out
5	RTO	one b m p in the open i c m in effect over
6	FSO	one b m p in the open i c m in effect out
7	FSO	message to observer, kilo alpha, high explosive, four rounds adjust fire, target number alpha bravo one zero zero zero, over
8	RTO	m t o kilo alpha four rounds target number alpha bravo one out
9	FSO	shot, over
10	RTO	shot out
11	FSO	splash, over
12	RTO	splash out
13	RTO	right five zero fire for effect over
14	FSO	right five zero fire for effect out
15	FSO	shot, over
16	RTO	shot out
17	FSO	rounds complete, over
18	RTO	rounds complete out
19	RTO	end of mission one b m p suppressed zero casualties over
20	FSO	end of mission one b m p suppressed zero casualties out

Figure 2 CFF dialogue with radiobot FSO

In the second phase (lines 7-12 of figure 2) the FSO takes dialogue initiative with a message to observer (MTO, line 7), which informs the FO team about details of the fire that will be sent: the units that will fire, the type of ammunition, number of rounds, the method of fire, and the target number. In lines 9 and 11 the FSO informs the team when the fire has been sent and when it is about to land. At each point, the RTO confirms the information.

After the resulting fire, the RTO regains initiative in the third phase (lines 13-20 of figure 2). Depending on the observed results, the mission may be closed, or the fire may be repeated with an adjustment in location or method of fire, in which case the dialogue repeats an abbreviated version of the first two phases. In this example (line 13), the FO requests the fire to be sent 50 meters to the right, and as a "fire for effect" bombardment, rather than the initial "adjust fire" targeting method. The FSO sends warnings for shot and completion of rounds (lines 15 and 17), and the RTO

closes the mission in line 19, describes the results and estimates casualties.

3. THE RADIOBOT-CFF SYSTEM

The core of our approach to system design was based on a detailed analysis of the CFF manual and a large number of transcripts from JFETS UTM training sessions with a human operator. This analysis led to a formal characterization of the information needed by a participant to represent and engage in this sort of dialogue, according to the information state approach to dialogue (Larsson and Traum, 2000). One of the key points is the definition of dialogue 'moves' and 'parameters' that convey the actions taken by participants in the course of a CFF dialogue. Engaging in dialogue can thus be reduced to the problems of deciding which moves and parameters are expressed by a given utterance (interpretation), how expressions affect the dialogue state and which moves and parameters should be produced in reply (dialogue management), and how to produce text for a given set of moves and parameters (generation). Figure 3 shows the dialogue moves and parameters from the first transmission in Figure 2, where the Identification dialogue move has as its parameters the call signs of the RTO and FSO, and the Warning Order dialogue move has as its parameters the method of fire requested and the method of target location.

IDENTIFICATION: steel one nine this is gator niner one
    fdc_id: steel one nine
    fo_id: gator nine one
WARNING ORDER: adjust fire polar
    method_of_fire: adjust fire
    method_of_location: polar

Figure 3 Dialogue moves and parameters

A total of 19 dialogue moves and 22 parameters were defined as the basic units for call for fire dialogue description (see Roque and Traum, 2006 for more detailed discussion).

The Radiobot-CFF system is made up of several pipelined components: Speech Recognizer, Interpreter, Dialogue Manager, and Generator.

The Speech Recognizer takes the audio signal of radio voice messages as input and produces text representations of what was said. It is implemented using the SONIC speech recognition system (Pellom, 2001) and was optimized for Radiobot-CFF with custom language and acoustic models derived from UTM training sessions and early test sessions of our system.

The Interpreter takes the output of the Speech Recognizer and determines what the utterance is trying to accomplish by identifying its dialogue moves and the parameters of those dialogue moves. The Interpreter uses a statistical approach, assigning the dialogue move and parameter to each word using a Conditional Random Field (Sha and Periera, 2003) tagger. The tagger looks at the statistical properties of word/label sequences to determine the dialogue move and parameter for each word, and was trained with 1,800 utterances hand-coded from our transcripts. The interpreter actually uses two taggers, one for dialogue moves and a separate one for parameters.

The Dialogue Manager uses the Information State approach (Larsson and Traum, 2000) to define relevant information on the status of the dialogue. The dialogue moves and parameters provided by the Interpreter are used to update the information state, which uses other rules to determine when to send messages to the simulator, and what kind of utterances to generate to the FO. The Dialogue Manager can be run in fully-automated, semi-automated, or manual mode, allowing the trainer to take over the session at any time.

The Generator uses templates to construct a text string from an information specification. In most cases the output is sent to the user in pre-recorded sound clips, although a speech synthesizer can be used in cases where there is no sound clip available.

Finally, mission information is sent to the FireSim XXI simulator, which realistically models fires and munitions for military analysis, and communicates with the UTM graphic and audio simulation to present those results to the observer team.

4. METHODS OF EVALUATION

There were several factors that influenced the overall goals and design of the evaluation criteria. Our evaluation goals include all of the following:

- Determination of the level of performance of the system as a whole
- Determination of the level of performance of specific components
- Determination of the effectiveness of the system for use in training in the UTM
- Determination of the user satisfaction, interacting with such technology
- Determination of approaches for improving the system

No single evaluation method could meet all of these evaluation goals. A typical method of dialogue system evaluation is to log system behavior and evaluate error rates per component. This has the advantage of being

objective and yielding precise quantitative results of the dialogue system's performance that are useful both for diagnosis for system improvement and for some degree of comparison across dialogue systems. Such an analysis does not measure the effectiveness of the system in the dialogue context – for example how the components are able to interact with each other and recover from errors, or how usable the system is. Objective measures of task success are necessary to evaluate the global effect of the dialogue system, though they risk conflating performance of the system, its integration with the simulator software, and the user's performance. In addition, though the main objective is to evaluate the system as system, the effect on the user's experience cannot be ignored. These considerations resulted in the combination of user questionnaires, objective performance measures and system component measures discussed below.

4.1. User questionnaires

User questionnaires covered three main areas: the participant's experience reflected by such measures as task difficulty and performance satisfaction; experience as RTO covering self ratings on performance, team member's performance as FO, and rating of dialogues with the FSO; and experience as FO self-rating and rating of team member's performance as RTO.

The Experience section of the questionnaire covered several factors of the subjects' general experience in the UTM, and were coded on a 1-5 scale, where 1= very low, 3= average, and 5= very high. Questions ranged over the degree of physical, mental and temporal demand the subjects experienced, degree of perceived performance success and satisfaction, and degree of frustration experienced.

The second section covered a team evaluation of the subject's experience as RTO. On a scale of 1-10, subjects were asked to rate their own overall performance as RTO, including specific performance ratings for adherence to correct CFF protocol and spoken fluency over the radio. They also rated their teammate's overall performance as FO.

The third section asked participants to rate, from their experience as RTO only, a number of factors covering their dialogue with the FSO (either human or radiobot, depending on the condition). Again on a scale of 1-10, subjects were asked how well they could understand the FSO, how well they thought the FSO understood them, the FSO's adherence to correct CFF protocol, spoken fluency and naturalness. Finally, they were asked if the FSOs performance or input affected their performance as RTO and, if so, to rate the affect from strongly negative to strongly positive.

The final section of the questionnaire asked participants to answer several of the questions above, but from the perspective of their experience as FO. These included an overall rating of their performance as FO, a rating of their teammate's performance as RTO, and whether (and to what degree) the FSO's performance affected their performance as FO.

4.2. Objective performance measures

The radiobot's performance was also evaluated on several objective mission performance measures. A mission was considered completed based on the user's initiative in sending an end of mission call. Most missions consist of several fire calls. To measure relative performance, we used three factors: time to fire, task completion rate, and accuracy.

Time to fire was measured in seconds for the initiating call of a mission only, as subsequent calls follow an abbreviated procedure, with some variations that were not directly comparable. To isolate system performance from user variation, time to fire was measured from the end of the user's first warning order radio transmission to the simulated fire.

Task completion rate was based on the number of unique warning orders initiated by the subject. Any warning orders subsequently cancelled by the subject on their own initiative (e.g. to revise their coordinates) were discounted.

Accuracy rate was taken from the total fires completed. To distinguish system performance from subject performance, a fire was considered accurate if sent to the location requested by the subject (regardless of the actual accuracy of the subject's target location).

4.3. Dialogue system component measures

To evaluate system component performance, we performed an analysis of session logs and human transcription and coded dialogue behavior to provide scores for the performance of the speech recognition, interpreter, and dialogue manager. The scores for each were averaged per session.

Speech recognition output was compared to hand transcribed utterances and was measured by two methods. The standard method, Word Error Rate (WER), is the ratio of mistakes to total correct words. We also included results in terms of F-score (the harmonic mean of Precision and Recall) for more straightforward comparison with the other components.

The Speech Interpreter was evaluated separately but in the same manner for its two components, dialogue moves and dialogue parameters. Speech recognizer results from the evaluation sessions were hand-coded with correct move and parameter values, then compared to the Interpreter's session output to yield a combined measure for the aggregate performance of Speech Recognizer + Interpreter (SI scores). The Interpreter's performance was also independently evaluated by obtaining interpreter results from the transcribed session utterances (I scores).

There is no standard metric for dialogue manager evaluation. We proposed a method for evaluation of information-state dialogue managers by calculating individual information state component F-scores between human judgements of the component and system values for each stage in the dialogue (Roque et al 2006a). We can also produce scores based on actual speech recognition and interpreter input (SID scores) as well as correct input (D score).

4.4. Dialogue generation analysis

Finally, to evaluate the resulting dialogue in performance, we analyzed the transcribed output of the Radiobot dialogue across fully automated sessions. Measures included the number of transmissions, the rate of response, the proportion of radiobot request for repair, and the proportion of correct responses.

5. EVALUATION PROCEDURE

The Radiobot-CFF evaluation was carried out in three phases: a preliminary evaluation, and two final evaluation sessions. The preliminary evaluation was conducted over two days in November 2005, with regular classes training in the UTM. Each team performed 2-4 calls for fire, and completed a questionnaire. While regular students were our ideal test case, we found that the objective of carrying out a well controlled study conflicted to some degree with the classroom needs of rotating a large number of students through the entire CFF training process. After the November test we also substantially refined the user questionnaire to more accurately reflect the experiences of the subjects in their respective roles as FO and RTO in evaluating both the dialogues with the FSO and their own performance. These revisions shaped the final evaluation, which was conducted in two sessions in January and February 2006.

The subjects for the final evaluation were volunteers, drawn primarily from two courses of training. This resulted in a fairly equal balance of two experience groups: the first were soldiers highly experienced in calls for fire, with substantial classroom and field training and, in most cases, real field experience. The second group ranged in experience from some classroom CFF training to complete novices in the domain, though all participants were soldiers experienced with standard army radio call procedures. Participants were given a group orientation prior to the experiment, in which they were given an overview of CFF procedures, answered demographic questionnaires, and signed up for test group time slots. Each team consisted of two participants, one from the highly experienced group, the other from the novice group.

There were three conditions that made up our evaluation:

- Fully Automated Condition: the radiobot acts as FSO, receiving and sending verbal transmissions with the RTO, and sends mission information to the simulator, without human operator intervention.
- Semi-Automated Condition: the radiobot dialogues with RTO and sends missions as above, but at each stage the information is displayed in a form which an operator may review and correct before submitting.
- Control Condition: a human acts as FSO, sending and receiving information from the RTO, while an operator enters mission information in a form and submits to the simulator.

Each participant attempted 2 missions (one grid and one polar mission) as FO, and 2 missions as RTO. Since we had more session time available than participants, some participants were run through multiple sessions in different teams. These participants were tracked, and care was taken to distribute their sessions across test conditions and randomize the order in which they were experienced. Likewise, we sought a balanced distribution based on experience and demographic information across each condition. After each test, participants filled out the questionnaire covering their experience.

6. RESULTS

We give results from several different approaches to the data below. User questionnaire data covers both of the final evaluation sessions; performance measures and dialogue system performance scores cover only the final February sessions.

6.1. User questionnaires

Questionnaire responses below include both January and February final evaluation dates. There were a total of 10 subjects in human sessions, 17 in semi-automated and 20 in fully automated.

As part of reviewing their experiences as RTO, participants were asked to rate their dialogue interaction

with the FSO, rating on a scale of 1-10 the following questions:

- Q1: How well could you understand the FSO?
- Q2: How well do you think the FSO understood you?
- Q3: How would you rate the FSO's adherence to correct Call for Fire protocol?
- Q4: How would you rate the FSO's spoken fluency on the radio?

The results are shown in table 1.

Table 1 Median rating of FSO dialogue

	Human	Semi	Auto
Q1	9	8	8
Q2	9	8	7.5
Q3	8.5	8	7.5
Q4	9	8	7.5

While the main objective of the radiobot is to allow for greater flexibility for the instructor and operators, it may only be considered successful if it doesn't significantly interfere with the trainee's experience and task success. As a measure of this, we asked participants to rate both their own and their teammate's performance in each role. The combined score is an average rating of both team members (self and other ratings) for each participant. RTO ratings are shown in table 2.

Table 2: Median RTO performance by condition

Rating	Human	Semi	Auto
Self	8	8	8.5
Other	9	9	8
Combined	8.5	8	8.25

The scores are quite comparable, with some variation across conditions, with again a slight preference for the human condition. The opposite trend holds for the FO ratings, however, in table 3, where performance with both radiobot conditions is slightly higher than with the human condition.

Table 3 Median FO performance by condition

Rating	Human	Semi	Auto
Self	8	9	8
Other	8	9	9
Combined	7.25	8.5	8.5

As another measure of the radiobot's effect on the participants' performance, they were asked if they felt the FSO's performance affected their own performance as RTO and FO and, if so, to rate the effect on a scale from 1-10, where 1= strongly negative and 10= strongly positive. Table 4 shows these results and the percentage of response indicating some effect on performance.

Table 4: Median Reported Effect on User Performance

	Human	Semi	Auto
RTO	6	5	6
% Response	30%	17.6%	35%
FO	4	5	5
% Response	10%	29.4%	40%

The reported affect on the RTO was nearly equal for the human and automated conditions, both in percent response and rating, with the semi-automated slightly lower. The reported affect on the FO, on the other hand was more noticeable given the higher response rate in both radiobot conditions, but also had a slightly positive rating over the human condition, which might be compared to the FO results from table 3 as well. In both cases, the radiobot conditions seem to have compared well to the human training condition, and met the goal of not significantly interfering in the trainees' performances.

6.2. Objective performance measures

Objective performance measures were calculated for the final February evaluation sessions only. The total number of missions for each condition, and performance per each condition, are shown in table 5.

Table 5 Mission performance by condition

	Human	Semi	Auto
Missions	11	17	21
Number of Fires	32	39	63
Fires per mission	2.9	2.3	3
Time to Fire	106.2	139.4	104.3
Task Completion	100%	97.5%	85.5%
Accuracy Rate	100%	97.4%	91.5%

The average time to fire for the fully automated condition was quite good, matching and slightly exceeding that of the human conditions. The semi-automated condition was approximately 40% slower on average, which largely reflected the delay from hand editing and verifying mission information and responses.

Task completion rate was quite good with the semi-automated condition, somewhat lower with the automated condition. Closer analysis revealed that the majority of the problems in the automated sessions appeared to be

Table 6 Dialogue generation performance across automated sessions

Session	System transmissions	Acks req	% Acks	Repair Requests	Correct responses	Flawless Responses	Flawless transmissions
W1-2	27	12	100%	8%	92%	58%	82%
W3-1	26	14	100%	14%	93%	50%	73%
T2-2	15	8	88%	0	71%	71%	87%
T4-2	21	13	85%	0	91%	46%	71%
T5-2	67	39	97%	11%	76%	53%	70%
T6-1	29	18	89%	0	75%	50%	66%
T6-2	13	6	100%	0	100%	83%	92%
T7-2	26	12	100%	0	92%	75%	89%
T9-1	29	18	83%	27%	87%	53%	72%
T9-2	22	12	92%	9%	100%	55%	77%
Median Scores	26	12.5	93.5%	4%	91.5%	54%	75%

due to integration issues between the main components (the radiobot dialogue manager, firesim, and UTM software), many of which have subsequently been fixed.

Of completed fires, the accuracy rate was again a bit lower in the fully automated condition. In the majority of cases, the error was due to the speech recognizer misinterpreting a digit from a grid location, or an additional add or adjust to the location.

6.3. Dialogue system measures

Dialogue component measures were calculated from the automated and semi-automated sessions from the February evaluation data. ASR performance had an average WER of 9.7% and an F-score of 0.93 across sessions.

The Interpreter alone (I score) had an overall F-score of 0.98 for dialogue moves and 0.98 for classifying dialogue parameters. When combined with Speech Recognition output (SI score), the Interpreter components achieved an overall F-score of 0.95 for processing dialogue moves, and an F-score of 0.93 for processing dialogue parameters.

The information state of the Dialogue Manager was hand coded and evaluated across the automated sessions per individual state component. There were a total of 22 components tracking the state the dialogue, and some variation in the results across these. The median score per component was .93 with corrected Interpreter input, and .82 with raw session input (see Roque et al 2006a for further detail).

6.4. Dialogue generation analysis

Table 6 shows the detailed results of our analysis of the system's dialogue output. The first column gives the total number of Radiobot transmissions during the user session, which gives a rough indication of the session length (recall this is not only a factor of the radiobot's performance, but also the number of adjustments made by the subjects). The second column shows the number of acknowledgments required of the system, while the third column shows the actual rate of system response. An acknowledgment was considered any system utterance responding to a user utterance that required some response. This includes all of the 'initiating' utterances of the RTO discussed in section 2, as well as any other requests for information. The median response rate was quite good, at 93.5%.

The rate of the radiobot's repair requests (e.g. 'Say again') is given in the fourth data column. This partially complements the rate of response, in that a request for repair is counted as an acknowledgment. Although there was some variation across sessions, the median rate of 4% is again quite good.

The final three columns give an indication of the quality of the radiobot's utterances. Columns 5 and 6 pertain only to radiobot transmissions that are responses to RTO utterances; Column 7 includes all radiobot transmissions. As responses depend on the RTO's transmitted information, and reflect the aggregate processing of the speech recognizer, classifier and dialogue manager, we expect the error rate to be higher than for other components. Even so, the median rate of correct responses was again quite high, at 91.5%. A response was considered correct if it conveyed all necessary semantic information for the given task to be completed, and occurred in the appropriate place in the dialogue.

We also applied a much stricter measure in calculating 'flawless' transmissions. A flawless transmission, in addition to being semantically correct, contained no errors in word output or protocol. Thus only 54% of the radiobot's responses but 75% of its total

transmissions could be considered flawless. Most of the errors under this measure were quite minor and do not affect the ultimate scenario performance, which is measured by the correctness rate of 91.5%. As they affect the sense of naturalness of the dialogue however, they should be corrected in further work. The errors fell into roughly three categories: errors of protocol (particularly a reversed ordering of left-right and add-drop adjustments), misrecognition of information that was not mission critical, and replication of noise from speech recognition input. The first two problems could be fairly easily corrected by added dialogue output constraints and additional training on more data. While noise in the output based on speech recognition will present a problem in any dialogue system, a combination of further training for improved recognition and additional constraints on the output string could improve those errors considerably.

CONCLUSIONS

Results of our evaluation across a variety of measures are encouraging. While there is still room for improvement compared to human-level performance, even this first version of the system performed well --- in many cases achieving over 90% performance level, which is sufficient to allow reduced human intervention for training exercises. Further goals for the improvement of the system will include a closer analysis of dialogue to evaluate domain specific dialogue appropriateness and protocol success in generation, as well as further investigation into more robust methods for error-handling. We are additionally performing linguistic analysis of human-human vs human-machine call for fire Dialogues (Martinovski and Vaswani 2006).

The potential impact on the warfighter of the further development and utilization of Radiobot technology should be apparent. Although simulated training may not replace the need for live training, the resources and expense of the latter often limit the trainee's exposure to real conditions. Simulations offer a useful supplemental resource, and the use of a radiobot in training simulations could enhance the efficiency of training, both by easing the load on the trainer while allowing multiple training simulations to run concurrently. Though our testbed for the radiobot was CFF training, the basic radiobot technology could be usefully expanded into numerous other training domains.

ACKNOWLEDGMENTS

We would like to thank the following people and organizations from Fort Sill, Oklahoma for their efforts on this project: the Depth & Simultaneous Attack Battle Lab, Techrizon, and Janet Sutton of the Army Research Laboratory. This work has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

Department of the Army, 1991: Tactics, techniques, and procedures for observed fire. Technical Report FM 6-30, Department of the Army.

Larsson, S. and D. Traum, 2000: Information state and dialogue management in the TRINDI dialogue move engine toolkit, *Natural Language Engineering*, **6**, Special Issue on Spoken Dialogue System Engineering, 323-340.

Martinovski, B., and A. Vaswani, 2006: Activity-based dialogue analysis as evaluation method, *Interspeech-06 Satellite Workshop Dialogue on Dialogues - Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*, September 17th, 2006.

Pellom, B., 2001: Sonic: The university of Colorado continuous speech recognizer. Technical Report TRCSLR-2001-01, University of Colorado.

Roque, A. and D. Traum, 2006: An information state-based dialogue manager for call for fire dialogues, *7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia, July 15-16.

Roque, A., H. Ai, and D. Traum, 2006: Evaluation of an information state-based dialogue manager.

Roque, A., A. Leuski, V. Rangarajan, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum, 2006: Radiobot-CFF: A spoken dialogue system for military training, *9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, Pittsburgh, PA, September 17-21, 2006.

Sha, F. and Pereira, F., 2003: Shallow parsing with conditional random fields, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*, 1, 134-141.